This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2856625, IEEE/ACM Transactions on Audio, Speech, and Language Processing

Cross-Domain and Semi-Supervised Named Entity Recognition in Chinese Social Media: A Unified Model

Jingjing Xu, Hangfeng He, Xu Sun, Xuancheng Ren and Sujian Li

***Abstract*—Named Entity Recognition (NER) in Chinese social media is an important but challenging task because Chinese social media language is informal and noisy. Most previous methods on NER focus on in-domain supervised learning, which is limited by scarce annotated data in social media. In this paper, we present that sufficient corpora in formal domains and massive unannotated text can be combined to improve NER performance in social media. We propose a unified model which can learn from out-of-domain corpora and in-domain unannotated text. The unified model is composed of two parts. One is for cross-domain learning and the other is for semi-supervised learning. Cross-domain learning can learn out-of-domain information based on domain similarity. Semi-supervised learning can learn in-domain unannotated information by self-training. Experimental results show that our unified model yields a 9.57% improvement over strong baselines and achieves the state-of-the-art performance[1].**

***Index Terms*—Named Entity Recognition, Chinese Social Media, Cross-Domain Learning, Semi-Supervised Learning.**

I. Introduction

NAMED entities are phrases that contain the names of persons, organizations, and locations. Identifying these entities in text is one of the basic tasks in natural language processing (NLP). The task is commonly known as Named Entity Recognition (NER). NER is very useful for many high-level tasks, such as information extraction and entity linking.

A long line of work [2], [3] focuses on formal text, e.g., news. Recently, with the great development of social media, a lot of researchers began to explore NER in social media [4]–[6]. NER in social media is challenging since social media text adopts more flexible and informal language usages than traditional formal text. Despite the recent progress in narrowing the performance gap between formal and social media domains in English [6], challenges remain in solving NER in Chinese social media.

Previous work mainly uses Conditional Random Field (CRF) or Structured Perceptron (SP) [7]–[11] to deal with NER in English or Chinese. For instance, Lin and Wu [9] used a linear-chain CRF with spelling features and phrase cluster features extracted from the web data. Ling and Weld [12] showed that the syntactic-level features from dependency structures in a CRF-based model improved NER performance. Recently, neural networks have achieved promising performance. For instance, Collobert et al. [13] used a CNN over a sequence of word embeddings and a CRF layer on top of the CNN output. Huang et al. [14] presented a similar model but used LSTM and hand-crafted spelling features. However, these methods rely on supervised learning that is limited by scarce annotated data.

In this paper, we propose a unified model that learns knowledge from both cross-domain annotated datasets and in-domain unannotated text to improve NER in Chinese social media. We refer to annotated datasets in formal domains as out-of-domain corpora and massive raw text in social media as in-domain unannotated text. In general, the proposed model is composed of two parts, a cross-domain learning module and a semi-supervised learning module.

The cross-domain learning module learns knowledge based on domain similarity. The core idea is to compute the similarities between out-of-domain training data and social media data. These similarities are used to adjust learning rates for out-of-domain sentences. Out-of-domain sentences with high similarities will influence training much more than sentences with low similarities. Furthermore, to avoid the distribution bias between out-of-domain and in-domain corpora, we introduce a novel similarity decay mechanism to our cross-domain learning

This work is a substantial extension of the paper presented at AAAI 2017 [1].

Jingjing Xu, Xu Sun, Xuancheng Ren and Sujian Li are with the MOE Key Laboratory of Computational Linguistics, Peking University, Beijing 100871, China (e-mail: jingjingxu@pku.edu.cn; xusun@pku.edu.cn; renxc@pku.edu.cn; lisujian@pku.edu.cn).

Hangfeng He is with the Departmeent of Computer and Information Science, University of Pennsylvania, Philadelphia, PA 19104, USA (e-mail: hangfeng@seas.upenn.edu).

[1]The code is released in https://github.com/lancopku/ChineseNER

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2856625, IEEE/ACM Transactions on Audio, Speech, and Language Processing

module, rather than use a fixed similarity weight to adjust the learning rate during training. As training epochs increase, the decay mechanism will adaptively decrease similarity weights for out-of-domain sentences.

The semi-supervised learning module learns in-domain knowledge by self-training. Previous work focuses on learning word representations from unannotated corpora [15], [16]. For example, Peng and Dredze [16] jointly trained models on NER and word segmentation tasks to learn word segmentation representations. Unlike previous work, we propose a confidence-based learning method, which explores in-domain unannotated text by self-training. To build training pairs, predictions made by the trained model before each epoch are regarded as the gold tags for unannotated text. We use confidence to evaluate the correctness of training pairs and adjust the learning rate for each training pair in unannotated corpora. The confidence is calculated by a confidence learning function. To reduce confidence calculation cost on massive unannotated data, we introduce a sentence ranking mechanism that chooses top k unannotated sentences based on the similarity between each unannotated sentence and an in-domain corpus.

Our main contributions are as follows:

- We propose a unified model that leverages both out-of-domain annotated datasets and in-domain unannotated text to improve NER in Chinese social media.
- We propose a cross-domain learning module with a decay mechanism that learns knowledge from out-of-domain datasets based on domain similarity.
- We design a semi-supervised learning module with a sentence ranking mechanism that learns knowledge from in-domain unannotated text by self-training.

II. Proposal

In this section, we will describe the basic model structure, the cross-domain learning module, the semi-supervised learning module, and the unified model in sequence.

A. BiLSTM-MMNN

Following previous work, we use a bidirectional LSTM as our basic model. To build a structured output, we also add the transition probability and max margin networks [17]. We refer to this basic model as BiLSTM-MMNN. We use character and position embeddings in the basic model.

Following the work of [17], the structured margin loss $\Delta(y, \bar{y})$ is defined as:

$$\Delta(y, \bar{y}) = \sum_{i=1}^{n} \kappa \mathbf{1}\{y_j \neq \bar{y_j}\} \tag{1}$$

where κ is a discount rate. $\mathbf{1}\{\cdot\}$ outputs 1 when the input is $True$ and outputs 0 otherwise. $\bar{y}$ is the predicted tag sequence. n is the length of gold tag sequence y. The loss is proportional to the number of characters with incorrect tags . For an input x, we search for the sequence with the highest score $\hat{y}$ as

$$\hat{y} = \underset{\bar{y} \in Y(x)}{\operatorname{argmax}}\, s(x, \bar{y}, \theta) \tag{2}$$

where $s(x, \bar{y}, \theta)$ represents the score of $\bar{y}$. $Y(x)$ is the set that contains all tag sequences. The score of gold sequence should be larger up to a margin than other sequences $\bar{y} \in Y(x)$:

$$s(x, \hat{y}, \theta) \geq s(x, \bar{y}, \theta) + \Delta(\hat{y}, \bar{y}) \tag{3}$$

To build a structured output, we also consider the transition probability in our network. The score of a tag sequence is

$$s(x, y, \theta) = \sum_{i=1}^{n} (A_{t_{i-1}t_i} + f_\Lambda(t_i|x)) \tag{4}$$

where n is the length of the tag sequence y. $A_{t_{i-1}t_i}$ represents the transition probability from tag t_{i-1} to t_i. $f_\Lambda(t_i|x)$ represents the probability of tag t_i. Λ represent the set of parameters in the basic model.

Character and Position Embeddings: Peng and Dredze [18] explored different kinds of embeddings for NER in Chinese social media: word, character, and character-positional embeddings. It first segmented NER data, and then learned word embeddings. For character embeddings, it directly learned embeddings in the training corpus. For character-positional embeddings, it was based on a character version but also considered the position of the character in a word. Experimental results showed that the character-positional embeddings achieved the best results. Following their work, we add segmentation information into our model, which combines the character and its position in a word together.

B. Similarity-Based Cross-Domain Learning with a Decay Mechanism

The motivation of our cross-domain module is based on the fact that if we directly use out-of-domain training sentences, the performance usually drops due to the distribution bias between in-domain and out-of-domain corpora. Thus, we first compute the similarities between

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2856625, IEEE/ACM Transactions on Audio, Speech, and Language Processing

out-of-domain and social media sentences, and then use the similarities to control the learning rates. The sentences with high similarity weights will influence training much more than the sentences with low similarity weights do.

We also find that fixing similarity weights during the whole training will result in several problems. First, it is difficult for models to learn in-domain data distribution because too many out-of-domain sentences are still involved into training during the later stage of training. Second, keeping similarity weights fixed without a decay mechanism would influence model convergence. To address these problems, we propose a novel similarity decay mechanism to decrease similarity weights dynamically during training.

The core idea is to use different learning rates for different training sentences. For out-of-domain sentences, the learning rate is controlled by the similarity weight. At epoch t, the learning rate for out-of-domain sentence x is computed as

$$\begin{aligned} LR(x,t) &= \alpha \cdot similar(x,I,t) \\ &= \alpha \cdot func(x,I) \cdot decay(x,t) \end{aligned} \tag{5}$$

where $similar(x,I,t)$ is the similarity weight between an out-of-domain sentence x and an in-domain corpus I at epoch t. α represents the initial learning rate for social media corpus I. $func(x,I)$ outputs the similarity between an out-of-domain sentence x and a social media corpus I. $decay(x,t)$ decides the decay speed of similarity for out-of-domain sentence x at epoch t.

Similarity Function: We design three different similarity functions.

Cross Entropy Function: As in paper [19], we use the cross entropy between out-of-domain sentences and an in-domain language model. This function reflects to what extent an input sentence matches the target data distribution. If a sentence follows the target data distribution, the language model will assign low entropy to it, and vice versa.

The detailed calculation process is:

$$\begin{aligned} &func(x,I) \\ &= C \cdot (-n^{-1}\log_2(\prod_{i=1}^{n} P(x_i|x_0,...,x_{i-1})))^{-1} \end{aligned} \tag{6}$$

where C is a hyper-parameter which is used to tune the magnitude of similarity. n is the length of sentence x. $P(x_i|x_0,...,x_{i-1})$ is the output probability produced by a language model trained on the in-domain corpus I.

Gaussian RBF Kernel Function: The detailed calculation is:

$$func(x,I) = C \cdot \exp(-\frac{\|v_x - v_I\|}{2\sigma^2}) \tag{7}$$

where C is a hyper-parameter used to tune the magnitude of similarity. σ is a hyper-parameter and used to adjust the variance. v_x is a vector representation for sentence x and v_I is a vector representation for social media dataset I. Sentence vector v_x is the mean of character-positional embeddings. Corpus vector v_I is the mean of sentence vectors. Character-positional embeddings are trained by word2vec [20].

Polynomial Kernel Function: The detailed calculation of polynomial kernel function is:

$$func(x,I) = C \cdot \frac{< v_x, v_I >^d}{\|v_x\|^d . \|v_I\|^d} \tag{8}$$

where C is a hyper-parameter and used to tune the magnitude of similarity. v_x is a vector representation for sentence x and v_I is a vector representation for social media dataset I. The calculation of v_I and v_x is same with that in Gaussian RBF kernel function. If $d = 1$, the polynomial kernel function can be written as $\frac{1}{C}cos\theta$, where θ represents the angle between v_x and v_I in Euclidean space.

Similarity Decay Function: The decay function further adjusts the weight of all out-of-domain data in the training process. In general, these data should matter less as the training proceeds, while some particular sentences could still be useful for the training. Hence, we design two decay terms in our similarity decay function. One is a global decay term, which decides the global decay speed for all out-of-domain sentences. The other is a local decay term, which is responsible for maintaining the importance of certain sentences. For sentence x at epoch t, the similarity decay function $decay(x,t)$ is calculated as

$$\begin{aligned} &decay(x,t) \\ &= \beta^t \frac{1}{n} \sum_{i=1}^{n} \frac{Unique(y_i^{t-(m-1)},...,y_i^{t-1},y_i^t)}{|m|} \end{aligned} \tag{9}$$

where β^t is the global decay term at epoch t that gradually decreases sentence weights, and $Unique(y_i^{t-(m-1)},...,y_i^{t-1},y_i^t)$ is the number of different predictions in the latest m epochs for word x_i.

The global decay term is responsible for decreasing the weight of out-of-domain data as the training proceeds. The motivation comes from that if the weights of all out-of-domain data keep fixed in the whole training process, it is hard for the model to fit the target data distribution because the distribution of out-of-domain

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2856625, IEEE/ACM Transactions on Audio, Speech, and Language Processing

dataset is different from that of the target dataset. Therefore, the global decay term is proposed to solve this problem by gradually reducing similarity weights as the training proceeds. In the early stage of training, all out-of-domain data are involved for learning new knowledge outside the small target dataset. In the later stage of training, the model learns more from target domain data to ensure that it fits the target data distribution.

We introduce the local decay term to assign different out-of-domain sentences with different decay speeds. The main idea is that we increase decay speeds for "useless" sentences while decreasing decay speeds for "useful" sentences. Given an out-of-domain sentence, if predictions given by the model always change in the latest m iterations, it suggests that the model is not confident to make predictions on this sentence. Therefore, it becomes useful for the model to learn knowledge for accurately identifying entities in this sentence, and we decrease its decay speed. In contrast, if predictions do not change in the latest m epochs, we regard that the model is very confident and it is hard for gold annotations to change model predictions. To accelerate convergence speed, we increase the value of local decay term for such sentences. Since sharing the same global decay term for all out-of-domain data is over simplified, the local decay term makes the decay function more flexible.

C. Confidence-Based Semi-Supervised Learning with Sentence Ranking

Since annotating data needs a lot of efforts, how to make use of unannotated text to learn social media knowledge increasingly attracts researchers' attention. Traditional semi-supervised methods first build training pairs on unannotated text and then select the most confident training pairs. Following this framework, we design a semi-supervised learning module, which is composed of two parts, confidence learning and unannotated sentence ranking.

Confidence Learning Function: The confidence learning function is based on decision boundary. Generally, models always adopt tag sequences with the highest scores as prediction results. We consider the margin between the tag sequence with the highest score and the tag sequence with the second highest score as the decision boundary. The confidence is defined as

$$confid(x) = \frac{\hat{y}(x) - y_{2nd}(x)}{\hat{y}(x)} \tag{10}$$

where the confidence is decided by the decision boundary between $\hat{y}(x)$ and $y_{2nd}(x)$. $\hat{y}(x)$ is the tag sequence with the highest score and $y_{2nd}(x)$ is the tag sequence with the second highest score.

For an unannotated sentence x, the confidence learning function first considers the tag sequence with the highest score as

$$\hat{y}(x) = \underset{\bar{y} \in Y(x)}{\operatorname{argmax}}\, s(x, \bar{y}, \theta) \tag{11}$$

where $Y(x)$ is the set of all candidate predictions. $\hat{y}(x)$ is the tag sequence with the highest score and $s(x, \bar{y}, \theta)$ is the output score of sentence $\bar{y}$ with parameter θ.

The confidence learning function also considers the tag sequence with the second highest score as

$$y_{2nd}(x) = \underset{\bar{y} \in Y(x) \text{ and } \bar{y} \neq \hat{y}}{\operatorname{argmax}}\, s(x, \bar{y}, \theta) \tag{12}$$

where $Y(x)$ is the set of all candidate predictions. $\hat{y}(x)$ is the tag sequence with the highest score and $s(x, \bar{y}, \theta)$ is the output score of sentence $\bar{y}$ with parameter θ.

We compute the confidence weights before every epoch. Since predictions are made by the model in training, the confidence weight changes adaptively in different epochs.

Unannotated Sentence Ranking: First, it is hard for a model to give out-of-domain unannotated data accurate predictions and too many low-quality pairs (unannotated sentence - predicted label) bring a negative influence on model performance. Second, making predictions for massive unannotated sentences costs too much time before each epoch. Motivated by those factors, we introduce a sentence ranking method which first chooses top-k sentences based on the similarity between unannotated data and an in-domain corpus. The similarity of unannotated sentence x and in-domain corpus I is the reciprocal of cross entropy of sentence x according to the output probability of the language model. This function reflects how likely an input matches the target data distribution. If a sentence follows the target data distribution, the language model will give it low entropy. The detailed selection process is shown as follows. We first train a basic language model based on in-domain text and calculate a confidence weight for unannotated sentence $x = x_1, x_2, ..., x_i, ..., x_N$ as

$$\begin{aligned} &p(x, I) \\ &= \frac{1}{Z} \cdot (-n^{-1} \log_2(\prod_{i=1}^{n} P(x_i | x_0, ..., x_{i-1})))^{-1} \end{aligned} \tag{13}$$

where n is the length of sentence x. $P(x_i|x_0, ..., x_{i-1})$ is the output probability produced by the language model trained on in-domain corpus I. Z is the normalization term and calculated as

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2856625, IEEE/ACM Transactions on Audio, Speech, and Language Processing

$$Z = \sum_{x \in X} (-n^{-1} \log_2 (\prod_{i=1}^{n} P(x_i | x_0, ..., x_{i-1})))^{-1} \quad (14)$$

where X is the set of all unannotated sentences.

Then, we rank all unannotated sentences based on $p(x, I)$ and choose top k unannotated sentences. Only the selected sentences can be used to train the model with confidence weights. We choose the most similar sentences rather than the most diverse sentences because it is hard for a model trained on target data to give accurate predictions for all unannotated sentences. The unannotated dataset used in our paper is sampled from the Internet and contains many out-of-domain sentences, which cause low prediction accuracy. The resulting low-quality prediction pairs (unannotated sentence - predicted label) bring many noises to training, leading to a lower performance. To avoid the negative influence of low-quality pairs, we choose the most similar unannotated sentences that are more likely to get accurate predictions to augment the training set.

For an unannotated sentence x, the learning rate $\alpha_t(x)$ at epoch t is calculated as

$$LR(x, t) = \alpha \cdot confid(x, t) \quad (15)$$

where α is the initial learning rate for all in-domain sentences. $confid(x, t)$ evaluates the confidence of sentence x at epoch t.

D. Unified Model

For a training sentence x, the learning rate at epoch t is defined as

$$LR(x, t) = \alpha \cdot weight(x, t) \quad (16)$$

where $weight(x, t)$ is the function for computing relative weight, and α is the initial learning rate. The detailed calculation of $weight(x, t)$ is

$$weight(x, t) = \begin{cases} 1.0 & x \in S_{in\_domain} \\ similar(x, I, t) & x \in S_{cross\_domain} \\ confid(x, t) & x \in S_{unannotated} \end{cases}$$

where S_{in_domain} represents the set of all in-domain annotated sentences, S_{cross_domain} represents the set of all out-of-domain annotated sentences, $S_{unannotated}$ represents the set of all in-domain unannotated sentences, $similar(x, I, t)$ outputs the similarity weight between an out-of-domain x and a social media corpus I at epoch t and $confid(x, t)$ outputs the confidence weight of unannotated sentence x at epoch t. For each social media sentence, the default weight is set to 1.

Given a training sentence x associated with the weight function $weight(x, t)$, the model parameter θ is updated as

$$\theta = \theta - LR(x, t) \cdot \nabla_\theta(x, y, \theta) \quad (17)$$

where $\nabla_\theta(x, y, \theta)$ is the gradient of θ with respect to the loss function.

III. Experiments

To demonstrate the effectiveness of our proposed model, we design comprehensive experiments on a Chinese social media dataset. The experimental settings and results are described in this section.

A. Datasets

Following Dredze [16], [18], we use the same corpus for Chinese social media NER. The corpus is constructed based on Sina Weibo[2]. The dataset contains four types of entities in total, including ORG (organization), GPE (geo-political), PER (person), and LOC (location) entities. Each entity has two sub-types, named and nominal mention. This dataset not only contains annotated sentences, but also provides massive unannotated text. The details of dataset are shown in Table I. We first segment unannotated text by a commonly-used Chinese word segmentation system: Jieba[3].

We use a corpus from the sixth SIGHAN Workshop on Chinese language Processing as our cross-domain dataset. The difference between the SIGHAN corpus and the social media corpus lies in that the SIGHAN dataset lacks GPE entities and only has one sub-type: named mention. The details of SIGHAN corpus are shown in Table II.

B. Baselines

We design two baselines to compare with our proposed model. We first train the basic model on the social media dataset without out-of-domain datasets and unannotated text. Next, we make use of the out-of-domain datasets and design the second baseline, which pre-trains the basic model on out-of-domain datasets and then trains it on the social media dataset. We refer to these two baseline methods as "BiLSTM-MMNN" and "BiLSTM-MMNN + All Data Merge".

[2]One of the most popular social network in China.

[3]https://github.com/fxsjy/jieba.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2856625, IEEE/ACM Transactions on Audio, Speech, and Language Processing

Table I
DETAILS OF WEIBO CORPUS.

	Named Mention Entity	**Nominal Mention Entity**
Train set	957	898
Development set	153	226
Test set	211	198
unannotated Text	112,971,734 Weibo messages	

Table II
DETAILS OF SIGHAN CORPUS.

Entity Type	**Train Set**	**Test Set**
Location	18522	3658
Organization	10261	2185
Person	9028	1864
Total	37811	7707

C. Settings

We first use word2vec [20] to pretrain word embeddings. Following the work of Mao [24], we design a feature table which uses bigram features:

$$C_n C_{n+1} (n = -2, -1, 0, 1) \quad and \quad C_{-1} C_1$$

We use the window approach [13] to extract high level features from word features. The model is trained by stochastic gradient descent with $L2$ regularization.

We set the window size in feature extraction to 5, the dimension of word embeddings and feature embeddings to 100, the dimension of hidden vectors to 100, the discount κ in margin loss to 0.2, the initial learning rate α to 0.1, the global decay term β to 0.90, and the magnitude tuning constant C in three similarity functions to 1.

D. Results

Table III shows the results of our proposed models in terms of F1-score of Named Mention Entity (NAM) and F1-score of Nominal Mention Entity (NOM). We also include overall micro F1-score and out-of-vocabulary (OOV) recall.

Comparing two baseline systems, we find that even with the help of out-of-domain datasets, "BiLSTM-MMNN + All Data Merge" still achieves the lowest results. It can be attributed to the problem of distribution bias. The out-of-domain dataset only contains NAMs. Too many NAMs result in label bias and bring a bad influence on results of NOM. By comparing our proposed methods with "BiLSTM-MMNN", we find that the cross-domain learning module mainly improves results in NAM F1-score and semi-supervised learning module mainly improves results in NOM F1-score. Finally, the unified model yields a 9.57% improvement over BiLSTM-MMNN in terms of micro F1-score.

Furthermore, we notice that with the large recall improvement achieved by the proposed method, the precision exhibits different degrees of decline. The lower precision of the proposed method is mainly because it predicts significantly more entities than the baselines. For NAMs, the proposed method improves recall from 30.80 to 50.23. It demonstrates that the proposed method is capable of predicting much more NAMs. Normally, with the large increase of recall, it is hard to keep the original precision. But the proposed method only brings a slight reduction of precision for NAMs, from 71.42 to 70.19. For NOMs, although the decline of precision is more obvious than that in NAMs, the improvement of recall is also almost the two times as large as the decline of precision, thus bringing a better overall score.

Table IV compares our unified model with the state-of-the-art models. Our unified model achieves the best results in terms of micro F1-score. Peng and Dredze [18], [21] corrected the dataset and updated results in [23][4]. We refer to the updated results as Peng and Dredze (2017a) and Peng and Dredze (2017b). Table IV shows that the unified model beats all state-of-the-art methods.

Cross-Domain Learning: We first evaluate three similarity functions: cross entropy, Gaussian RBF kernel and polynomial kernel. The detailed hyper parameter settings are shown as follows. The magnitude tuning constant is set as $C = 1$. We use a trigram language model in cross entropy function. We set σ to 1 in Gaussian RBF kernel function and d to 1 in polynomial kernel function. As shown in Table VI, the polynomial kernel function achieves the best overall performance. The Gaussian function computes the absolute distance between a sentence vector and a corpus vector to evaluate how well a sentence matches a corpus. In real-world tasks, we find that the absolute distance for most sentences is too small, which reduces the distinguishing

[4]Due to the dataset was constructed by Amazon Mechanical Turk, the final annotations were generated by merging labels from multiple different Turkers. This inevitably leads to inconsistencies and errors. Thus, Peng and Dredze corrected the dataset and undated the results in [23].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2856625, IEEE/ACM Transactions on Audio, Speech, and Language Processing

Table III
RESULTS OF OUR PROPOSED METHOD ON THE TEST SET.

Models	Named Mention Entity			Nominal Mention Entity				
	Precision	Recall	F1	Precision	Recall	F1	OOV	Overall
BiLSTM-MMNN	67.30	33.17	44.44	70.45	46.96	56.36	16.95	50.21
+All Data Merge	71.42	30.80	43.04	35.41	25.75	29.82	19.99	36.64
Cross-Domain Learning (Our Proposal)	69.23	51.18	58.85	68.70	51.01	58.30	29.56	58.70
Semi-Supervised Learning (Our Proposal)	55.30	34.59	42.56	65.31	57.07	60.91	19.56	51.44
Unified Model (Our Proposal)	70.19	50.23	58.56	65.69	57.07	61.08	33.04	**59.78**

Table IV
COMPARISONS WITH STATE-OF-THE-ART METHODS.

Models	Named Mention Entity			Nominal Mention Entity				
	Precision	Recall	F1	Precision	Recall	F1	OOV	Overall
Peng and Dredze (2015) [18]	57.98	35.57	44.09	63.84	29.45	40.38	*	42.70
Peng and Dredze (2016) [21]	63.33	39.18	48.41	58.59	37.42	45.67	*	47.38
He and Sun (2017a) [22]	66.93	40.67	50.60	66.46	53.57	59.32	20.96	54.82
He and Sun (2017b) [1]	61.68	48.82	54.50	74.13	53.54	62.17	28.70	58.23
Peng and Dredze (2017a) [23]	74.78	39.81	51.96	71.92	53.03	61.05	*	56.05
Peng and Dredze (2017b) [23]	66.67	47.22	55.28	74.48	54.55	62.97	*	58.99
Unified Model (Our Proposal)	70.19	50.23	58.56	65.69	57.07	61.08	33.04	**59.78**

Table V
COMPARISONS OF USED RESOURCES BETWEEN OUR METHOD WITH STATE-OF-THE-ART METHODS. "TOTAL SIZE" DENOTES THE TOTAL NUMBER OF SENTENCES IN USED TRAINING SETS.

Models	Used Resources	Total Size	Overall Performance
Peng and Dredze (2015) [18]	Weibo corpus	1350	42.70
Peng and Dredze (2016) [21]	Weibo corpus + word segmentation data	45313	47.38
He and Sun (2017a) [22]	Weibo corpus + cross-domain data + unannotated data	34532	54.82
He and Sun (2017b) [1]	Weibo corpus+ cross-domain data + unannotated data	34532	58.23
Peng and Dredze (2017a) [23]	Weibo corpus	1350	56.05
Peng and Dredze (2017b) [23]	Weibo corpus+ word segmentation data	45313	58.99
Unified Model (our proposal)	Weibo corpus + cross-domain data + unannotated data	34532	**59.78**

Table VI
RESULTS OF THE CROSS-DOMAIN LEARNING MODULE ON THE TEST SET.

Models	Named Mention Entity			Nominal Mention Entity				
	Precision	Recall	F1	Precision	Recall	F1	OOV	Overall
BiLSTM-MMNN	67.30	33.17	44.44	70.45	46.96	56.36	16.95	50.21
Cross Entropy Function	52.46	40.28	45.57	*	*	*	22.60	23.51
Gaussian RBF Kernel Function	67.33	47.86	55.95	89.28	12.62	22.12	22.17	39.57
Polynomial Kernel Function	62.42	48.81	54.78	80.85	19.19	31.02	21.73	43.28
+ Similarity Decay Function	67.49	51.18	58.22	77.27	17.17	28.09	23.47	43.63
+ Post processing	69.23	51.18	58.85	68.70	51.01	58.30	29.56	**58.70**

ability. In contrast, the polynomial kernel function can be written as $\frac{1}{C}cos\theta$ when $d = 1$, where θ represents the angle between a sentence vector and a corpus vector in Euclidean space. The advantage of this function is that the distance is regularized by the vector length. Therefore, it evaluates relative distance, rather than absolute distance, and the similarity weights given by this function are more accurate. Based on the experimental results, we adopt the polynomial kernel function as our similarity function.

Table VI shows that the proposed similarity decay approach brings almost 4% improvement in NAMs. To show how the performance is affected by the choice of different decay terms, we report NAM F-score based on different global decay values in Figure 1. It is important to note that $\beta = 1$ represents that only the local decay term works in such case. The dotted line denotes the results of cross-domain learning without the similarity

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2856625, IEEE/ACM Transactions on Audio, Speech, and Language Processing

Table VII
COMPARISONS WITH THE RESULTS IN THE EARLIER CONFERENCE PAPER.

Models	Named Mention Entity			Nominal Mention Entity				
	Precision	Recall	F1	Precision	Recall	F1	OOV	Overall
He and Sun (2017b) [1]	61.68	48.82	54.50	74.13	53.54	62.17	28.70	58.23
+ Cross Domain Learning (Our proposal)	67.49	51.18	58.85	68.70	51.01	58.30	29.56	58.70
+ Semi-Supervised Learning (Our proposal)	70.19	50.23	58.56	65.69	57.07	61.08	33.04	**59.78**

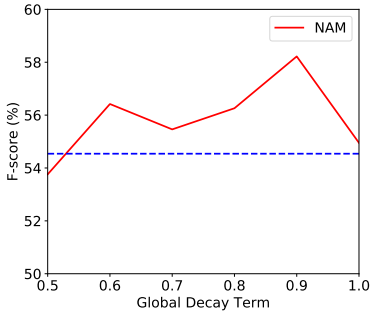

Figure 1. Results of the similarity decay function with different global decay terms. The dotted line denotes the result of cross-domain learning without the similarity decay mechanism.

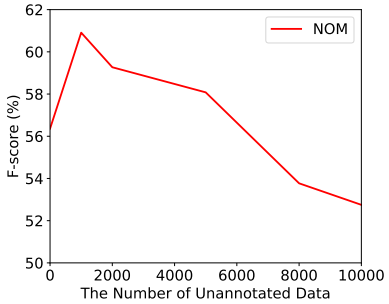

Figure 2. Results of the semi-supervised learning module with different amounts of unannotated data. Our semi-supervised learning method achieves the best results with 2000 unannotated data.

decay function. By comparing the method without the similarity decay function and the method only with the local decay term ($\beta = 1$), we find that the latter achieves better results, which proves the effectiveness of the local decay term.

Moreover, from Figure 1, we also find that the model with $\beta = 0.9$ achieves the best performance, and the models with too small or too large the value of β perform worse. If the global decay term is too small, the weights of out-of-domain sentences would decay so fast that the model can not learn enough out-of-domain knowledge. If the global decay term is too large, the weights of out-of-domain sentences would decay slowly, and thus it will lead to distribution bias during the later period of training.

We also notice that cross-domain learning deteriorates the results of NOM, as shown in Table VI. The size of SIGHAN dataset is almost 10 times of Weibo dataset and the SIGHAN dataset only contains NAMs. Too many NAMs result in label bias, thus compromising the performance of NOM. To solve this problem, we design a post processing method which combines the results of cross-domain learning and BiLSTM-MMNN. We keep the NOMs predicted by BiLSTM-MMNN and then combine it with the NAMs predicted by cross-domain learning.

Semi-Supervised Learning: We first select a subset of unannotated sentences based on a sentence ranking method. The pre-trained BiLSTM-MMNN is used to initialize parameters.

The results with different amounts of unannotated data are shown in Figure 2. We use "Overall F1-score" to represent micro F1-score between NAMs and NOMs. Since sentence ranking is based on the similarity between each unannotated sentence and an in-domain corpus. The most similar and important sentences are more likely to be chosen to train the model. Thus, even with the minimum subset of unannotated data, the proposed semi-supervised method also achieves the comparable results. In Figure 2, we can see that as the percentage of unannotated data increases, the performance becomes better at first. The method of using 1000 unannotated sentences achieves the best result. However, the performance gets worse with the further increase of out-of-domain sentences, because too many dissimilar sentences are involved into training.

Unified Model: Our unified model improves the performance of NAMs, but compromises the performance of NOM, like cross-domain module does. To address this problem, we adopt the same post processing method to combine the results of cross-domain learning and semi-supervised learning. The process keeps the NOMs predicted by semi-supervised learning and then combines it with the NAMs predicted by cross-domain learning. Table III shows that the unified model outperforms the baselines with 9.57% improvement in overall score. Furthermore, Table IV shows that the unified model beats all state-of-the-art methods. Table V shows that

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2856625, IEEE/ACM Transactions on Audio, Speech, and Language Processing

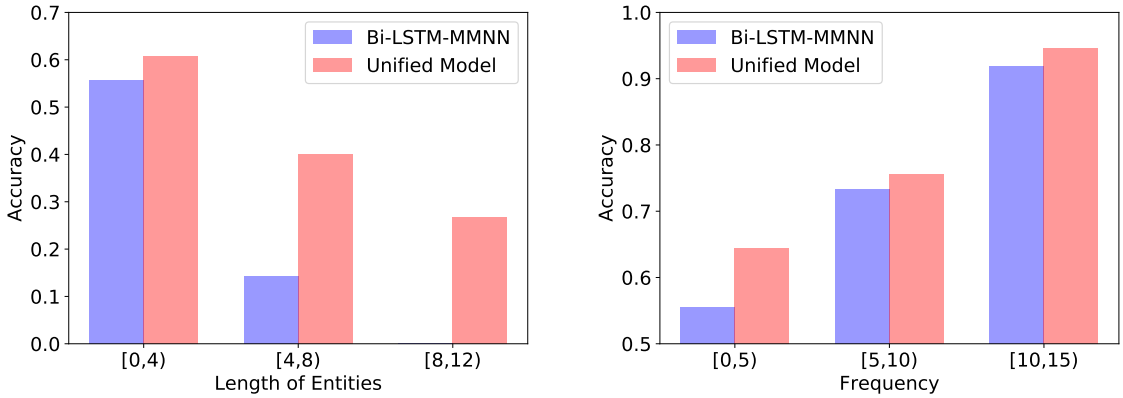

Figure 3. Results of the unified model and BILSTM-MMNN. Red bars represent the wrong predictions and purple bars represent the right predictions. As we can see, the basic model performs worse at long entities or entities with low frequency in training dataset.

Table VIII
EXAMPLES GENERATED BY THE CROSS-DOMAIN LEARNING METHOD AND THE BASELINE METHOD: BILSTM-MMNN.

Gold	叶葆(PER-NAM):全球时尚财运滚滚而来钱.
BiLSTM-MMNN	叶葆:全球时尚财运滚滚而来钱.
Cross-Domain Learning (Proposal)	叶葆(PER-NAM):全球时尚财运滚滚而来钱.
Gold	日中午,宋同志(PER-NAM)抵达汉口站(LOC-NAM)转动车回家.
BiLSTM-MMNN	日中午,宋同志抵达汉口站转动车回家.
Cross-Domain Learning (Proposal)	日中午,宋同志抵达汉口站(LOC-NAM)转动车回家.
Gold	发表了一篇转载博文转载龙门石窟(LOC-NAM) 潜溪寺(LOC-NAM).
BiLSTM-MMNN	发表了一篇转载博文转载龙门石窟潜溪寺.
Cross-Domain Learning (Proposal)	发表了一篇转载博文转载龙门石窟(LOC-NAM)潜溪寺.

Table IX
EXAMPLES GENERATED BY THE SEMI-SUPERVISED LEARNING METHOD AND THE BASELINE METHOD: BILSTM-MMNN.

Gold	@一句心情签名:老师(PER-NOM),你这么有才,学校(ORG-NOM)知道么?
BiLSTM-MMNN	@一句心情签名:老师(PER-NOM),你这么有才,学校(PER-NOM)知道么?
Semi-Supervised Learning (Proposal)	@一句心情签名:老师(PER-NOM),你这么有才,学校(ORG-NOM)知道么?
Gold	你真是了解宅女(PER-NOM)的好宅男(PER-NOM啊!
BiLSTM-MMNN	你真是了解宅女的(PER-NOM)好宅男(PER-NOM)啊!
Semi-Supervised Learning (Proposal)	你真是了解宅女(PER-NOM)的好宅男(PER-NOM)啊!
Gold	哥(PER-NOM),新婚快乐呀.
BiLSTM-MMNN	哥,新婚快乐呀.
Semi-Supervised Learning (Proposal)	哥(PER-NOM),新婚快乐呀.

our method also achieves the best performance compared with the work of using the same resources. It shows the effectiveness of the proposed method that takes advantage of cross-domain and unannotated data. To analyze the improvements, we compare the unified model with Bi-LSTM-MMNN from the following two aspects: the length of entities and the frequency. From Figure 3, we find that the baseline performs worse when entities are long or entities do not appear in training data. In contrast, our proposed unified model performs better.

For a clear understanding about the proposed method in this paper, Table VII explicitly compares the performance achieved in the earlier conference paper and the performance achieved in this paper after introducing each of changes to the method. The newly proposed cross-domain learning module largely improves the results of NAMs whiling declining the results of NOMs, thus leading a slight overall improvement. With the newly proposed semi-supervised learning module that improves the results of NOMs, our method improves the results more obviously and outperforms the original method in the conference paper with 1.55 point in overall score.

IV. ERROR ANALYSIS

With the help of massive cross-domain corpora and unannotated social media text, the proposed model achieves large improvements compared with the strong baselines. However, as we can see from experimental

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2856625, IEEE/ACM Transactions on Audio, Speech, and Language Processing

Table X
EXAMPLES OF ERROR TYPES.

NO-CROSS	Gold	延参**(PER-NAM)** 法师品味人生如同走进一片山水,静静的呼吸,安静的欣赏,这就是生活.
	Prediction	**延参**法师品味人生如同走进一片山水,静静的呼吸,安静的欣赏,这就是生活.
CONTAIN	Gold	嘻嘻女孩纸们**(PER-NOM)**随时算算量量哦
	Prediction	嘻嘻女孩**(PER-NOM)**纸们随时算算量量哦
BE-CONTAINED	Gold	明显宾哥比我更需要好吧！@小月子**(PER-NAM)c**
	Prediction	明显宾哥比我更需要好吧！@小月子c**(PER-NAM)**
CROSS	Gold	以前还特意逃课去拍这些花花草草#早安西电**(ORG-NAM)**#
	Prediction	以前还特意逃课去拍这些花花草草#早安西**(PER-NAM)**电#

results, the best results in Chinese social media are still much lower than that in the news domain. The state-of-the-art result in social media is 59.78% and the state-of-art result in the SIGHAN dataset is 92.81%. Thus, we need to do error analysis to help us understand which factor contributes most the low results in Chinese social media. Moreover, the improvements achieved by the proposed methods also need to be explored in detail. In this paper, to do error analysis, we design six metrics as follows:

- Entity length.
- Sentence length.
- Five error types: CONTAIN[5], CONTAINED[6], SPLIT[7], CROSS[8], NO-CROSS[9]. The distribution of error types are shown in Figure 4.
- Frequency of entity in training dataset.
- OOV rate[10] in a sentence.
- OOV rate in an entity.

A. Cross-Domain Learning

By analyzing the results of the baseline and the cross-domain learning method, we list several key improvements achieved by the cross-domain learning method as follows:

- Cross-Domain learning significantly decreases NO-CROSS errors by improving the recall of NAMs.
- Cross-Domain learning improves the results of entities which do not appear in training data or appear infrequently.

To understand how the cross-domain learning method improves results, we show several generated examples in Table VIII. We can see that the cross-domain learning method mainly improves the recall rate of NAMs and correctly identifies the entities that are hard to be recognized by the baseline, such as "叶葆 (Yebao, a person name)", "汉口站 (Hankou Station, a location name)", and "龙门石窟 (Longmen Grottoes, a location name)". The common point of these entities lies in that they either do not appear in training data or appear infrequently. In general, cross-domain corpora provide abundant information about words and entities that do not appear in social media data. It is a good choice to use out-of-domain data to broaden the knowledge of models.

[5]Gold tags contain predicted tags.
[6]Gold tags are contained by prediction tags.
[7]There are gaps in predicted tags.
[8]Gold tags cross predicted tags.
[9]There is no same words between gold tags and predicted tags.
[10]we refer to out-of-vocabulary words as OOV.

B. Semi-Supervised Learning

By analyzing the results of the baseline and the semi-supervised learning method, we find that the semi-supervised learning method mainly improves NOM results. The reason is that the NOM (e.g., 'girl') is more common than the NAM (e.g., 'Jane') in unannotated sentences. Since unannotated sentences augment more NOM-related training pairs, the semi-supervised learning method mainly improves results on NOM, with 4.55% F-score. Table IX shows generated examples predicted by the semi-supervised learning method and the baseline method, Bi-LSTM-MMNN. We can observe that the baseline method sometimes identifies the infrequent entities wrong. With the help of unannotated data, the semi-supervised learning method addresses this problem to some extent by increasing the number of infrequent entities.

C. Unified Model

Since the proposed model leverages massive cross-domain corpora and unannotated social media text, it achieves large improvements compared to several strong baselines. However, the best results on Chinese social media achieved in our work are still much lower than that in the news domain. By analyzing our results in detail, we list the most serious problems which needs to be addressed as follows:

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2856625, IEEE/ACM Transactions on Audio, Speech, and Language Processing

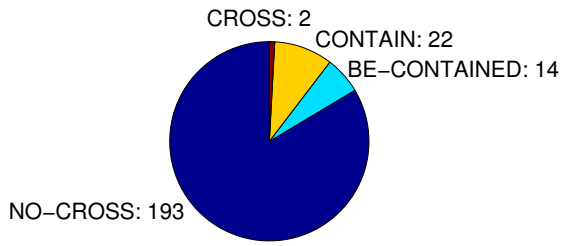

Figure 4. Error types of entities.

- NO-CROSS errors. Table X and Figure 4 show the common error types and related examples. From Figure 4, we can see that the percentage of NO-CROSS error is 83.55%, which covers most cases in wrong predictions. The problem of NO-CROSS error mainly comes from the low recall rate. Some methods for improving recall rate should be explored in the future, e.g. the re-sampling mechanism for better identifying infrequent entities by balancing training data.
- Infrequent entities in training dataset. It is hard to solve this problem only by the model itself. Additional knowledge bases are also required.
- Entities with OOV words or entities in sentences with OOV words. For this problem, we plan to improve the generalization ability of the model for better handling OOV words, like using character-level embeddings.

V. Background and Related Work

In this paper, we focus on how to leverage cross-domain annotated datasets and unannotated social media text with deep learning to improve Chinese social media NER. In this section, we first briefly review this task and then introduce the related work.

A. Chinese Social Media Named Entity Recognition

NER is a task which aims to recognize special entities in given text and to divide these entities into different category, such as person or location. The standard annotation guideline[11] defines five entity categories in total, such as location ('American'), titles ('Red and Black'), person ('John'), organizations ('APEC') and so on. Every single type also has three sub-types, such as pronominal phrase sub-type, nominal phrase sub-type, and name phrase sub-type. Traditional NER systems only consider name phrase. Unlike them, we explore two sub-types in this work, including name phrase entities and nominal phrase entities.

NER is a basic NLP task and there are a lot of related researches. One standard approach for NER is to regard the problem as a sequence labeling problem, where each word is assigned with a tag, indicating whether a word belongs to any named entity or appears outside of all entities. Previous approaches used sequence labeling models, such as hidden Markov models [25], maximum entropy Markov models [26], as well as conditional random fields [10].

While most research efforts exploited standard word-level features, more sophisticated features can also be used. Ling and Weld [12] showed that using syntactic-level features from dependency structures in a CRF-based model can improve NER performance.

More recently, neural networks have achieved promising results [13], [14]. Huang et al. [14] presented a new CRF-LSTM models but using hand-crafted spelling features. Collobert et al. [13] used CNN+CRF structure to handle this task.

These researches mainly focus on formal domains, e.g., news. However, NER in Chinese social media is a more difficult task. Lacking explicit word boundaries, which are helpful for Chinese NER, brings another challenge. How to improve this task is attracting more and more attention. For example, Peng and Dredze [18] explored three types of neural embeddings for representing Chinese text and trained embeddings simultaneously for NER and language modeling. Peng and Dredze [16] jointly trained NER and word segmentation to learn word segmentation representations in NER. In this paper, we use a confidence function to evaluate the training pairs in unannotated corpora. To reduce confidence calculation cost on massive unannotated data, we introduce a sentence ranking mechanism that chooses top k unannotated sentences based on the similarity between each unannotated sentence and an target domain corpus. There are some studies that use the top-k technique to choose unannotated example [27]–[30]. Different from these methods, we evaluates the similarity using a language model, rather than a self-trained model. The advantage is that it saves time since the calculation of similarity before each epoch is not necessary.

B. Cross-Domain Learning

First, not every target domain has enough annotated datasets since annotating data needs too much time and efforts. Second, training on cross-domain datasets may not bring significant improvements as one would expect due to the distribution bias. Third, for predicting domain-

[11]Entities V1.7, Linguistic Data Consortium, 2014

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2856625, IEEE/ACM Transactions on Audio, Speech, and Language Processing

unknown data, the performance usually drops if distribution bias is not considered in training models. Thus, there are a lot of researches which aim to take advantage of cross-domain datasets to achieve better performance in target domains [31], [32]. Lui and Baldwin [33] manually designed several domain-related features to help model learn domain information. Axelrod, He, and Gao [34] created pseudo examples in the target domain to learn target distributions. Bhatt, Semwal, and Roy [31] used another similarity function to compute task similarity. Wen [35] directly taught models to learn domain-specific knowledge.

C. Semi-Supervised Learning

Large unannotated text which is easily obtained can be used to improve the target domain performance. Thus, how to explore these corpora attracts a lot of researchers' attention.

A lot of semi-supervised learning methods were proposed to explore leverage unannotated text. These methods usually assign unannotated data with predicted labels, and the most confident examples are usually selected for training [36], [37]. Watson and Briscoe [38] used the method of self-training to select confident examples. Sarkar [39] and Maeireizo, Litman, and Hwa [40] used the method of co-training and adopted two classifiers to make predictions for unannotated datasets.

For semi-supervised learning in NER, previous work focuses on learning word representations from unannotated corpora [15], [16]. For example, Peng and Dredze [16] jointly trained models on NER and word segmentation tasks to learn word segmentation representations.

VI. Conclusions

In this paper, we propose a unified model for Chinese social media NER. The model can learn from both out-of-domain annotated datasets and in-domain unannotated text. We assign different learning rates for different kinds of data. For out-of-domain annotated sentences, the learning rate is adjusted by a cross-domain learning module. For unannotated sentences, the learning rate is adjusted by a semi-supervised learning module.

The cross-domain learning module contains two parts, a similarity function and a similarity decay function. The similarity function is used to evaluate the similarity between every cross-domain sentence and the in-domain dataset. The similarity decay function is used to dynamically decrease similarity weights during training to solve the problem of distribution bias.

The semi-supervised module leverages massive unannotated text by self-training. We first build training pairs based on predictions made by the model before each epoch and then use a confidence function to evaluate the correctness of training pairs. Since confidence calculation for massive unannotated text costs too much time, we propose a sentence ranking method to choose the most similar and important sentences.

In our experiments, the cross-domain learning module and semi-supervised learning module both achieve large improvements compared to the baseline systems in terms of NAM F1-score and NOM F1-score, respectively. The unified model yields a 9.57% improvement over a strong baseline and outperforms several state-of-the-art models. Furthermore, our detailed and targeted result analysis not only helps us understand the improvements of our proposed model but also points out the direction of future work.

Acknowledgements

This work is a substantial extension of the paper presented at AAAI 2017 [1]. We thank Nanyun Peng for providing the dataset. We thank the reviewers who provide thoughtful suggestions. This work was supported in part by National Natural Science Foundation of China (No. 61673028). Xu Sun is the corresponding author of this paper.

References

[1] H. He and X. Sun, "A unified model for cross-domain and semi-supervised named entity recognition in Chinese social media," in *AAAI*, 2017, pp. 3216–3222.

[2] A. McCallum and W. Li, "Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons," in *HLT-NAACL*, 2003, pp. 188–191.

[3] G. Jin and X. Chen, "The fourth international Chinese language processing bakeoff: Chinese word segmentation, named entity recognition and Chinese POS tagging," in *IJCNLP*, 2008, pp. 69–81.

[4] M. Dredze, P. McNamee, D. Rao, A. Gerber, and T. Finin, "Entity disambiguation for knowledge base population," in *COLING 2010*, 2010, pp. 277–285.

[5] C. Li and Y. Liu, "Improving named entity recognition in tweets via detecting non-standard words." in *ACL*, 2015, pp. 929–938.

[6] C. Cherry and H. Guo, "The unreasonable effectiveness of word representations for twitter named entity recognition." in *HLT-NAACL*, 2015, pp. 735–745.

[7] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *ICML*, 2001, pp. 282–289.

[8] N. Ye, W. S. Lee, H. L. Chieu, and D. Wu, "Conditional random fields with high-order features for sequence labeling," in *NIPS*, 2009, pp. 2196–2204.

[9] D. Lin and X. Wu, "Phrase clustering for discriminative learning," in *ACL*, 2009, pp. 1030–1038.

[10] X. Sun, "Structure regularization for structured prediction," in *NIPS*, 2014, pp. 2402–2410.
[11] X. Sun, T. Matsuzaki, D. Okanohara, and J. Tsujii, "Latent variable perceptron algorithm for structured classification," in *IJCAI*, 2009, pp. 1236–1242.
[12] X. Ling and D. S. Weld, "Fine-grained entity recognition." in *AAAI*, 2012.
[13] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. P. Kuksa, "Natural language processing (almost) from scratch," in *JMLR*, vol. 12, 2011, pp. 2493–2537.
[14] Z. Huang, W. Xu, and K. Yu, "Bidirectional LSTM-CRF models for sequence tagging," in *CoRR*, vol. abs/1508.01991, 2015.
[15] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," in *NAACL*, 2016, pp. 260–270.
[16] N. Peng and M. Dredze, "Improving named entity recognition for Chinese social media with word segmentation representation learning," in *ACL*, 2016.
[17] W. Pei, T. Ge, and B. Chang, "Max-margin tensor neural network for Chinese word segmentation." in *ACL*, 2014, pp. 293–303.
[18] N. Peng and M. Dredze, "Named entity recognition for Chinese social media with jointly trained embeddings," in *Proceedings of EMNLP*, 2015, pp. 548–554.
[19] A. Axelrod, X. He, and J. Gao, "Domain adaptation via pseudo in-domain data selection," in *EMNLP*, 2011, pp. 355–362.
[20] T. Mikolov and J. Dean, "Distributed representations of words and phrases and their compositionality," in *NIPS*, 2013.
[21] N. Peng and M. Dredze, "Improving named entity recognition for Chinese social media with word segmentation representation learning," in *ACL*, vol. 2, 2016, pp. 149–155.
[22] H. He and X. Sun, "F-score driven max margin neural network for named entity recognition in Chinese social media," in *EACL*, 2017, pp. 713–718.
[23] N. Peng and M. Dredze, "Supplementary results for named entity recognition on Chinese social media with an updated dataset," Tech. Rep., 2017. [Online]. Available: http://www.cs.jhu.edu/~npeng/papers/golden_horse_supplement.pdf
[24] X. Mao, Y. Dong, S. He, S. Bao, and H. Wang, "Chinese word segmentation and named entity recognition based on conditional random fields." in *IJCNLP*, 2008, pp. 90–93.
[25] G. Zhou and J. Su, "Named entity recognition using an HMM-based chunk tagger," in *ACL*, 2002, pp. 473–480.
[26] A. McCallum, D. Freitag, and F. C. Pereira, "Maximum entropy markov models for information extraction and segmentation." in *ICML*, vol. 17, 2000, pp. 591–598.
[27] W. Pan, X. Shen, A. Jiang, and R. P. Hebbel, "Semi-supervised learning via penalized mixture model with application to microarray sample classification," in *Bioinformatics*, vol. 22, no. 19, 2006, pp. 2388–2395.
[28] M. Chang, L. Ratinov, and D. Roth, "Guiding semi-supervision with constraint-driven learning," in *ACL*, 2007.
[29] T. Jebara, J. Wang, and S. Chang, "Graph construction and *b*-matching for semi-supervised learning," in *ICML*, 2009, pp. 441–448.
[30] Y. Cheng, W. Xu, Z. He, W. He, H. Wu, M. Sun, and Y. Liu, "Semi-supervised learning for neural machine translation," in *ACL*, 2016.
[31] H. S. Bhatt, D. Semwal, and S. Roy, "An iterative similarity based adaptation technique for cross domain text classification," in *CoNLL*, 2015.
[32] H. S. Bhatt, M. Sinha, and S. Roy, "Cross-domain text classification with multiple domains and disparate label sets." in *ACL*, 2016.
[33] M. Lui and T. Baldwin, "Cross-domain feature selection for language identification," in *IJCNLP*, 2011.
[34] A. Axelrod, X. He, and J. Gao, "Domain adaptation via pseudo in-domain data selection," in *EMNLP*, 2011, pp. 355–362.
[35] T.-H. Wen, M. Gasic, N. Mrksic, L. M. Rojas-Barahona, P.-H. Su, D. Vandyke, and S. Young, "Multi-domain neural network language generation for spoken dialogue systems," in *CoRR*, 2016.
[36] D. Kawahara and K. Uchimoto, "Learning reliability of parses for domain adaptation of dependency parsing." in *IJCNLP*, vol. 8, 2008.
[37] J. Yu, M. Elkaref, and B. Bohnet, "Domain adaptation for dependency parsing via self-training," in *IWPT*, 2015.
[38] R. Watson and T. Briscoe, "Adapting the rasp system for the conll07 domain-adaptation task." in *EMNLP-CoNLL*, 2007, pp. 1170–1174.
[39] A. Sarkar, "Applying co-training methods to statistical parsing," in *NAACL*, 2001, pp. 1–8.
[40] B. Maeireizo, D. Litman, and R. Hwa, "Co-training for predicting emotions with spoken dialogue data," in *ACL*, 2004, p. 28.

Jingjing Xu is a Ph.D. student, supervised by Prof. Xu Sun, at School of Electronics Engineering and Computer Science, Peking University. Her recent research interests include deep learning applied to natural language processing.

Hangfeng He is a first-year Ph.D. student in the Department of Computer and Information Science at the University of Pennsylvania. His research deals with machine learning and natural language processing. Currently, he works on indirect supervision for semantics representation, interpretability of deep learning models and machine reading comprehension.

Xu Sun is Associate Professor in Department of Computer Science, Peking University, since 2012. He got Ph.D from The University of Tokyo (2010), M.S. from Peking University (2007), and B.E. from Huazhong Univ. of Sci. & Tech. (2004). His research focuses on natural language processing and machine learning.

Xuancheng Ren is a Ph.D. candidate, supervised by Prof. Xu Sun, at School of Electronics Engineering and Computer Science, Peking University. He received the degree of Bachelor of Science in Computer Science from Peking University in 2017. His work focuses on machine learning for natural language processing, especially on deep learning methods.

Sujian Li received the Ph.D. degree from the Institute of Computing Technology, Chinese Acamedy of Sciences, Beijing, China, in 2002. She is currently an Associate Professor at the Institute of Computational Linguistics, Peking University, Beijing. Her main research interests include natural language processing, information extraction, and document summarization.